

What is claimed is:

- 1           1.       A system for grouping clusters of semantically scored documents,  
2 comprising:  
3           a scoring module determining a score assigned to at least one concept  
4 extracted from a plurality of documents based on at least one of a frequency of  
5 occurrence of the at least one concept within at least one such document, a  
6 concept weight, a structural weight, and a corpus weight; and  
7           a clustering module forming clusters of the documents by applying the  
8 score for the at least one concept to a best fit criterion for each such document.
- 1           2.       A system according to Claim 1, further comprising:  
2           the scoring module calculating the score as a function of a summation of  
3 at least one of the frequency of occurrence, the concept weight, the structural  
4 weight, and the corpus weight of the at least one concept.
- 1           3.       A system according to Claim 2, further comprising:  
2           a compression module compressing the score through logarithmic  
3 compression.
- 1           4.       A system according to Claim 1, further comprising:  
2           the scoring module calculating the concept weight as a function of a  
3 number of terms comprising the at least one concept.
- 1           5.       A system according to Claim 1, further comprising:  
2           the scoring module calculating the structural weight as a function of a  
3 location of the at least one concept within the at least one such document.
- 1           6.       A system according to Claim 1, further comprising:  
2           the scoring module calculating the corpus weight as a function of a  
3 reference count of the at least one concept over the plurality of documents.
- 1           7.       A system according to Claim 1, further comprising:

2           the scoring module forming the score assigned to the at least one concept  
3   to a normalized score vector for each such document, determining a similarity  
4   between the normalized score vector for each such document as an inner product  
5   of each normalized score vector, and applying the similarity to the best fit  
6   criterion.

1           8.     A system according to Claim 1, further comprising:  
2           the clustering module evaluating a set of candidate seed documents  
3   selected from the plurality of documents, identifying a set of seed documents by  
4   applying the score for the at least one concept to a best fit criterion for each such  
5   candidate seed document, and basing the best fit criterion on the score of each  
6   such seed document.

1           9.     A method for grouping clusters of semantically scored documents,  
2   comprising:  
3           determining a score assigned to at least one concept extracted from a  
4   plurality of documents based on at least one of a frequency of occurrence of the at  
5   least one concept within at least one such document, a concept weight, a structural  
6   weight, and a corpus weight; and  
7           forming clusters of the documents by applying the score for the at least  
8   one concept to a best fit criterion for each such document.

1           10.    A method according to Claim 9, further comprising:  
2           calculating the score as a function of a summation of at least one of the  
3   frequency of occurrence, the concept weight, the structural weight, and the corpus  
4   weight of the at least one concept.

1           11.    A method according to Claim 10, further comprising:  
2           compressing the score through logarithmic compression.

1           12.    A method according to Claim 9, further comprising:  
2           calculating the concept weight as a function of a number of terms  
3   comprising the at least one concept.

- 1           13.     A method according to Claim 9, further comprising:  
2           calculating the structural weight as a function of a location of the at least  
3 one concept within the at least one such document.
- 1           14.     A method according to Claim 9, further comprising:  
2           calculating the corpus weight as a function of a reference count of the at  
3 least one concept over the plurality of documents.
- 1           15.     A method according to Claim 9, further comprising:  
2           forming the score assigned to the at least one concept to a normalized  
3 score vector for each such document;  
4           determining a similarity between the normalized score vector for each  
5 such document as an inner product of each normalized score vector; and  
6           applying the similarity to the best fit criterion.
- 1           16.     A method according to Claim 9, further comprising:  
2           evaluating a set of candidate seed documents selected from the plurality of  
3 documents;  
4           identifying a set of seed documents by applying the score for the at least  
5 one concept to a best fit criterion for each such candidate seed document; and  
6           basing the best fit criterion on the score of each such seed document.
- 1           17.     A computer-readable storage medium holding code for performing  
2 the method of Claim 9.
- 1           18.     A system for providing efficient document scoring of concepts  
2 within a document set, comprising:  
3           a frequency module determining a frequency of occurrence of at least one  
4 concept within a document retrieved from the document set; and  
5           a concept weight module analyzing a concept weight reflecting a  
6 specificity of meaning for the at least one concept within the document;

7 a structural weight module analyzing a structural weight reflecting a  
8 degree of significance based on structural location within the document for the at  
9 least one concept;  
10 a corpus weight module analyzing a corpus weight inversely weighing a  
11 reference count of occurrences for the at least one concept within the document;  
12 and  
13 a scoring module evaluating a score associated with the at least one  
14 concept as a function of the frequency, concept weight, structural weight, and  
15 corpus weight.

1 19. A system according to Claim 18, further comprising:  
2 the scoring module evaluating the score substantially in accordance with  
3 the formula:

$$4 \quad S_i = \sum_{j \rightarrow n}^j f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

5 where  $S_i$  comprises the score,  $f_{ij}$  comprises the frequency,  $0 < cw_{ij} \leq 1$  comprises  
6 the concept weight,  $0 < sw_{ij} \leq 1$  comprises the structural weight, and  $0 < rw_{ij} \leq 1$   
7 comprises the corpus weight for occurrence  $j$  of concept  $i$ .

1 20. A system according to Claim 19, further comprising:  
2 the concept weight module evaluating the concept weight substantially in  
3 accordance with the formula:

$$4 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

5 where  $cw_{ij}$  comprises the concept weight and  $t_{ij}$  comprises a number of terms for  
6 occurrence  $j$  of each such concept  $i$ .

1 21. A system according to Claim 19, further comprising:  
2 the structural weight module evaluating the structural weight substantially  
3 in accordance with the formula:

$$4 \quad sw_{ij} = \begin{cases} 1.0, & \text{if}(j \approx \text{SUBJECT}) \\ 0.8, & \text{if}(j \approx \text{HEADING}) \\ 0.7, & \text{if}(j \approx \text{SUMMARY}) \\ 0.5 & \text{if}(j \approx \text{BODY}) \\ 0.1 & \text{if}(j \approx \text{SIGNATURE}) \end{cases}$$

5 where  $sw_{ij}$  comprises the structural weight for occurrence  $j$  of each such concept  $i$ .

1 22. A system according to Claim 19, further comprising:  
2 the corpus weight module evaluating the corpus weight substantially in  
3 accordance with the formula:

$$4 \quad rw_{ij} = \begin{cases} \left( \frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

5 where  $rw_{ij}$  comprises the corpus weight,  $r_{ij}$  comprises a reference count for  
6 occurrence  $j$  of each such concept  $i$ ,  $T$  comprises a total number of reference  
7 counts of documents in the document set, and  $M$  comprises a maximum reference  
8 count of documents in the document set.

1 23. A system according to Claim 19, further comprising:  
2 a compression module compressing the score substantially in accordance  
3 with the formula:

$$4 \quad S'_i = \log(S_i + 1)$$

5 where  $S'_i$  comprises the compressed score for each such concept  $i$ .

1 24. A system according to Claim 18, further comprising:  
2 a global stop concept vector cache maintaining concepts and terms; and  
3 a filtering module filtering selection of the at least one concept based on  
4 the concepts and terms maintained in the global stop concept vector cache.

1 25. A system according to Claim 18, further comprising:

2 a parsing module identifying terms within at least one document in the  
3 document set, and combining the identified terms into one or more of the  
4 concepts.

1 26. A system according to Claim 25, further comprising:  
2 the parsing module structuring each such identified term in the one or  
3 more concepts into canonical concepts comprising at least one of word root,  
4 character case, and word ordering.

1 27. A system according to Claim 25, wherein at least one of nouns,  
2 proper nouns and adjectives are included as terms.

1 28. A system according to Claim 18, further comprising:  
2 a plurality of candidate seed documents;  
3 a similarity module determining a similarity between each pair of a  
4 candidate seed document and a cluster center;  
5 a clustering module designating each such candidate seed document  
6 separated from substantially all cluster centers with such similarity being  
7 sufficiently distinct as a seed document, and grouping each such candidate seed  
8 document not being sufficiently distinct into a cluster with a nearest cluster  
9 center.

1 29. A system according to Claim 28, further comprising:  
2 a plurality of non-seed documents;  
3 the similarity module determining the similarity between each non-seed  
4 document and each cluster center; and  
5 the clustering module grouping each such non-seed document into a  
6 cluster having a best fit, subject to a minimum fit criterion.

1 30. A system according to Claim 29, further comprising:  
2 a normalized score vector for each document comprising the score  
3 associated with the at least one concept for each such concept occurring within  
4 the document; and

5 the similarity module determining the similarity as a function of the  
6 normalized score vector associated with the at least one concept for each such  
7 document.

1 31. A system according to Claim 30, further comprising:  
2 the similarity module calculating the similarity substantially in accordance  
3 with the formula:

$$4 \quad \cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

5 where  $\cos \sigma_{AB}$  comprises a similarity between a document  $A$  and a document  $B$ ,  
6  $\vec{S}_A$  comprises a score vector for document  $A$ , and  $\vec{S}_B$  comprises a score vector for  
7 document  $B$ .

1 32. A system according to Claim 29, further comprising:  
2 a dynamic threshold module determining a dynamic threshold for each  
3 cluster based on the similarities between each document in the cluster and a center  
4 of the cluster; and  
5 the similarity module identifying each outlier document having such a  
6 similarity outside the dynamic threshold.

1 33. A system according to Claim 32, further comprising:  
2 the clustering module grouping each such outlier document into a cluster  
3 having a best fit, subject to a minimum fit criterion and the dynamic threshold of  
4 the cluster.

1 34. A system according to Claim 32, wherein the dynamic threshold is  
2 determined based on the similarities of the documents in the cluster to the cluster  
3 center.

1 35. A method for providing efficient document scoring of concepts  
2 within a document set, comprising:

3 determining a frequency of occurrence of at least one concept within a  
4 document retrieved from the document set; and  
5 analyzing a concept weight reflecting a specificity of meaning for the at  
6 least one concept within the document;  
7 analyzing a structural weight reflecting a degree of significance based on  
8 structural location within the document for the at least one concept;  
9 analyzing a corpus weight inversely weighing a reference count of  
10 occurrences for the at least one concept within the document; and  
11 evaluating a score associated with the at least one concept as a function of  
12 the frequency, concept weight, structural weight, and corpus weight.

1 36. A method according to Claim 35, further comprising:  
2 evaluating the score substantially in accordance with the formula:

$$3 \quad S_i = \sum_{j=1 \rightarrow n}^j f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

4 where  $S_i$  comprises the score,  $f_{ij}$  comprises the frequency,  $0 < cw_{ij} \leq 1$  comprises  
5 the concept weight,  $0 < sw_{ij} \leq 1$  comprises the structural weight, and  $0 < rw_{ij} \leq 1$   
6 comprises the corpus weight for occurrence  $j$  of concept  $i$ .

1 37. A method according to Claim 36, further comprising:  
2 evaluating the concept weight substantially in accordance with the  
3 formula:

$$4 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

5 where  $cw_{ij}$  comprises the concept weight and  $t_{ij}$  comprises a number of terms for  
6 occurrence  $j$  of each such concept  $i$ .

1 38. A method according to Claim 36, further comprising:  
2 evaluating the structural weight substantially in accordance with the  
3 formula:



$$4 \quad sw_{ij} = \begin{cases} 1.0, & \text{if}(j \approx \textit{SUBJECT}) \\ 0.8, & \text{if}(j \approx \textit{HEADING}) \\ 0.7, & \text{if}(j \approx \textit{SUMMARY}) \\ 0.5 & \text{if}(j \approx \textit{BODY}) \\ 0.1 & \text{if}(j \approx \textit{SIGNATURE}) \end{cases}$$

5 where  $sw_{ij}$  comprises the structural weight for occurrence  $j$  of each such concept  $i$ .

1 39. A method according to Claim 36, further comprising:  
2 evaluating the corpus weight substantially in accordance with the formula:

$$3 \quad rw_{ij} = \begin{cases} \left( \frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

4 where  $rw_{ij}$  comprises the corpus weight,  $r_{ij}$  comprises a reference count for  
5 occurrence  $j$  of each such concept  $i$ ,  $T$  comprises a total number of reference  
6 counts of documents in the document set, and  $M$  comprises a maximum reference  
7 count of documents in the document set.

1 40. A method according to Claim 36, further comprising:  
2 compressing the score substantially in accordance with the formula:  
3  $S'_i = \log(S_i + 1)$

4 where  $S'_i$  comprises the compressed score for each such concept  $i$ .

1 41. A method according to Claim 35, further comprising:  
2 maintaining concepts and terms in a global stop concept vector cache; and  
3 filtering selection of the at least one concept based on the concepts and  
4 terms maintained in the global stop concept vector cache.

1 42. A method according to Claim 35, further comprising:  
2 identifying terms within at least one document in the document set; and  
3 combining the identified terms into one or more of the concepts.

1 43. A method according to Claim 42, further comprising:

2 structuring each such identified term in the one or more concepts into  
3 canonical concepts comprising at least one of word root, character case, and word  
4 ordering.

1 44. A method according to Claim 42, further comprising:  
2 including as terms at least one of nouns, proper nouns and adjectives.

1 45. A method according to Claim 35, further comprising:  
2 identifying a plurality of candidate seed documents;  
3 determining a similarity between each pair of a candidate seed document  
4 and a cluster center;  
5 designating each such candidate seed document separated from  
6 substantially all cluster centers with such similarity being sufficiently distinct as a  
7 seed document; and  
8 grouping each such candidate seed document not being sufficiently  
9 distinct into a cluster with a nearest cluster center.

1 46. A method according to Claim 45, further comprising:  
2 identifying a plurality of non-seed documents;  
3 determining the similarity between each non-seed document and each  
4 cluster center; and  
5 grouping each such non-seed document into a cluster with a best fit,  
6 subject to a minimum fit criterion.

1 47. A method according to Claim 46, further comprising:  
2 forming a normalized score vector for each document comprising the  
3 score associated with the at least one concept for each such concept occurring  
4 within the document; and  
5 determining the similarity as a function of the normalized score vector  
6 associated with the at least one concept for each such document.

1 48. A method according to Claim 47, further comprising:  
2 calculating the similarity substantially in accordance with the formula:

$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

where  $\cos \sigma_{AB}$  comprises a similarity between a document  $A$  and a document  $B$ ,  
 $\vec{S}_A$  comprises a score vector for document  $A$ , and  $\vec{S}_B$  comprises a score vector for  
document  $B$ .

49. A method according to Claim 46, further comprising:  
determining a dynamic threshold for each cluster based on the similarities  
between each document in the cluster and a center of the cluster; and  
identifying each outlier document having such a similarity outside the  
dynamic threshold.

50. A method according to Claim 49, further comprising:  
grouping each such outlier document into a cluster with a best fit, subject  
to a minimum fit criterion and the dynamic threshold of the cluster.

51. A method according to Claim 49, wherein the dynamic threshold is  
determined based on the similarities of the documents in the cluster to the cluster  
center.

52. A computer-readable storage medium holding code for performing  
the method of Claim 35.

53. An apparatus for providing efficient document scoring of concepts  
within a document set, comprising:  
means for determining a frequency of occurrence of at least one concept  
within a document retrieved from the document set; and  
means for analyzing a concept weight reflecting a specificity of meaning  
for the at least one concept within the document;  
means for analyzing a structural weight reflecting a degree of significance  
based on structural location within the document for the at least one concept;

- 9 means for analyzing a corpus weight inversely weighing a reference count  
10 of occurrences for the at least one concept within the document; and  
11 means for evaluating a score associated with the at least one concept as a  
12 function of the frequency, concept weight, structural weight, and corpus weight.